УДК 004.89

МОХАМЕД А. АБДЕЛААЛ, ст. группы 03-220
([1]Арабская нефтегазовая академия, [2]КФУ)
г. Абу-Даби

АЛГОРИТМ LISTBB И ЕГО МАТЕМАТИЧЕСКИЕ ОСНОВЫ

Алгоритм Listbb – это один из логических методов классификации, так как данная модель основана на методике построения бинарного решающего дерева по набору прецедентов [1, 2].

При решении задачи классификации всегда существует обучающая выборка (1):

$$X^l = (x_i; y_i)_{i=1}^l \subset X \times Y, y_i = y(x_i) \quad (1)$$

В формуле (1) множество Y – это конечное множество классов. Формализуем интуитивное понятие «логической закономерности».

Логическая закономерность – это некоторое условие с двумя значениями «0» и «1» (ложь и истина соответственно), которое описывается функцией [3]. Функция является отображением из множества объектов множества «0»/«1» и она удовлетворяет двум требованиям:

1. Требование интерпретируемости:
   - R(x) записывается на естественном языке [4];
   - R зависит от небольшого числа признаков (до 7), так как она не должна быть слишком сложной.
2. Требование информативности относительно одного из классов $y \in Y$:
   - $p_y(R) \neq \{x_i: R(x_i) = 1 \text{ и } y_i = y\} \to max;$
   - $n_y(R) \neq \{x_i: R(x_i) = 1 \text{ и } y_i \neq y\} \to min.$

Из второго условия следует, что оно записано более формализованным математическим языком [5].

Это позволяет задать возможность записи математического ограничения: необходимо требовать от предиката (2), чтобы он выдавал 1 преимущественно на объектах одного из классов.

$$\frac{p_y(R)}{P_y} \gg \frac{n_y(R)}{N_y} \quad (2)$$

Иными словами, необходимо определить число позитивных (то есть относящихся к классу «1») объектов класса y.

То есть «позитив» $p_y(R)$ – это число объектов выборки, на которых предикат возвращает единицу, и этот объект принадлежит нужному классу y.

А «негатив» $n_y(R)$ – это количество объектов выборки, на которых предикат возвращает единицу, но при этом объект не принадлежит к нужному нам классу (т.е. предикат ошибся) [6].

Таким образом, задачей предиката является выделить некоторое существенное подмножество объектов заданного класса и при этом не выделять объекты других классов (рисунок 1).

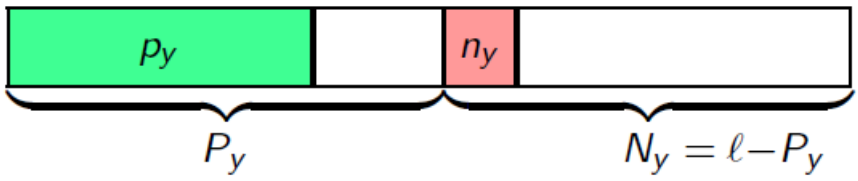

Рисунок 1. Задача выделения предиката

Из схематичного рисунка 1 и формул **(1)** и **(2)** очевидно, что задачей предиката является увеличить как можно больше объектов p_y и максимально сократить число объектов n_y.

Теперь из закономерностей необходимо построить алгоритм классификации. Для этого существует алгоритм индукции правил, который лежит в основе модели LISTBB (алгоритма синтеза бинарного решающего дерева, где LIST – списковое представление; Branching (B) – ветвление, а Boolean (B) – обозначает бинарную основу построения дерева решений).

Формализуем стратегию индукции правил:

1. Определить, какого вида закономерности необходимо найти (выбрать семейство правил для поиска закономерностей);
2. Сформулировать поисковый алгоритм, чтобы перебирать правила-кандидаты для поиска закономерностей;
3. Выяснить, по какому критерию информативности необходимо определять, является ли данное правило закономерностью или нет. То есть на данном шаге определяется критерий информативности;
4. Построить классификатор (3) из правил как из признаков.

$$a(x) = \arg max_{y \in Y} \sum_{j=1}^{n_y} w_{yj} R_{yj}(x) \quad (3)$$

В формуле (3) происходит взвешенное голосование о том, правила какого класса покрыли данный объект w_{yj} в большинстве.

Если объект будет покрыт 10 правилами одного класса и всего 3 правилами другого класса, тогда он будет отнесен к первому классу, так как он более надежно покрывается закономерностями первого класса [1].

В модели Listbb в качестве семейства правил часто используются три разных типа условий:

1. Пороговое условие (4). Их может быть несколько или только одно. Когда условие только одно, то пороговое условие называется «решающий пень», «decision stump»:

$$R(x) = [f_j(x) \le a_j] \text{ или } [a_j \le f_j(x) \le b_j] \quad (4)$$

где a_j, b_j – это пороги функций обучения.

2. Конъюнкция пороговых условий **(5)**:

$$R(x) = \wedge_{j \in J} \left[a_j \le f_j(x) \le b_j \right] (5)$$

где $\wedge_{j \in J}-$ это подмножество признаков для интерпретируемого содержательного понимания экспертом предметной области.

3. Синдром (выполнение не менее d условий из $|J|$) (6):

$$R(x) = \left[\sum_{j \in Y} \left[a_j \le f_j(x) \le b_j \right] \ge d \right] (6)$$

Отметим, что все параметры $J, a_j, b_j, d -$ это настраиваемые по обучающей выборке параметры. Они настраиваются путем решения оптимизационной задачи по критерию информативности [7, 8].

Список литературы:

1. Яцевич, М. Ю. Формирование модели сильного искусственного интеллекта на основе принципа "Congruit universa" для решения геомеханической задачи методом межскважинного сейсмоакустического просвечивания / М. Ю. Яцевич, П. А. Пылов, А. В. Дягилева // Вестник научного центра по безопасности работ в угольной промышленности. – 2022. – № 4. – С. 14-19. – EDN JOZUTB.
2. Свидетельство о государственной регистрации программы для ЭВМ № 2024611038 Российская Федерация. Sliding Window Predictor LLM : № 2023689243 : заявл. 25.12.2023 : опубл. 17.01.2024 / П. А. Пылов. – EDN AGNEAO.
3. Свидетельство о государственной регистрации программы для ЭВМ № 2023669862 Российская Федерация. Insight IQ : № 2023669164 : заявл. 19.09.2023 : опубл. 21.09.2023 / Р. В. Майтак. – EDN YCLTHP.
4. Свидетельство о государственной регистрации программы для ЭВМ № 2024611019 Российская Федерация. Multi-view generation of 3D objects based on diffusion model : № 2024610024 : заявл. 02.01.2024 : опубл. 17.01.2024 / П. А. Пылов. – EDN CQQJRO.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023680103 Российская Федерация. Cognitive Solution : № 2023669189 : заявл. 19.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QEMFJA.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2024610491 Российская Федерация. Modified Switch Transformer : № 2023689285 : заявл. 25.12.2023 : опубл. 11.01.2024 / П. А. Пылов. – EDN LYISWR.
7. Свидетельство о государственной регистрации программы для ЭВМ № 2023668767 Российская Федерация. Модель NLTK для мультизадачной обработки текста на русском языке : № 2023667990 : заявл. 01.09.2023 : опубл. 04.09.2023 / Р. В. Майтак. – EDN WVAUTY.
8. I. Isaev, S. Dolenko. Group Determination of Parameters and Training with Noise Addition: Joint Application to Improve the Resilience of the Neural Network Solution of a Model Inverse Problem to Noise in Data. Advances in Intelligent Systems and Computing, 2019, V.848, pp. 138-144. Springer, Cham. DOI: 10.1007/978-3-319-99316-4_18 (дата обращения: 21.08.2023).